DOCUMENT RESUME

ED 481 117 TM 035 280

AUTHOR	Gandal, Matthew
TITLE	Multiple Choices: How Will States Fill in the Blanks in Their Testing Systems?
PUB DATE	2002-02-00
NOTE	20p.; Paper prepared for the conference "Will No Child Truly Be Left Behind? The Challenges of Making This Law Work?" (February 13, 2002).
PUB TYPE	Opinion Papers (120) -- Speeches/Meeting Papers (150)
EDRS PRICE	EDRS Price MF01/PC01 Plus Postage.
DESCRIPTORS	*Elementary Secondary Education; *Federal Legislation; Standardized Tests; *Standards; *State Programs; *Test Construction; Test Use; *Testing Programs
IDENTIFIERS	*No Child Left Behind Act 2001

ABSTRACT

The new Elementary and Secondary Education Act amendments require states to begin annual testing in grades 3 through 8 in reading and mathematics by the 2005-2006 school year. Only 16 states currently have grade-by-grade tests in reading and mathematics, and only 9 of those have tests aligned with their standards, as the law requires. The rest of the states must fill in the blanks, and to do so, they need to consider whether states, the market, and the public are ready. Four scenarios suggest the different approaches the states might take: (1) purchasing ready-made tests; (2) letting districts use their own local tests; (3) developing new, customized tests; and (4) pooling resources with other states to develop new assessments. If past experience is a guide, the federal government may lay down markers and use the bully pulpit, but state leaders will have to address these testing problems themselves. (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

T.L. Pache

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Multiple Choices

How will states fill in the blanks in their testing systems?

ED 481 117

Matthew Gandal
Achieve

Draft – Please do not cite or quote without permission of the author.

TM035280

Prepared for

Will No Child Truly Be Left Behind?
The Challenges of Making This Law Work

A Conference Sponsored by the Thomas B. Fordham Foundation

February 13, 2002

BEST COPY AVAILABLE

Multiple Choices
How will states fill in the blanks in their testing systems?

By Matthew Gandal, Achieve

If someone had told me a couple of years ago that, over the next few years, every state was going to institute a grade-by-grade testing system, I would have laughed and thought that person was out of touch with reality and, frankly, politically naïve. Most states hadn't even established academic standards in each grade, let alone tests, and some were experiencing significant resistance from educators in the few grades where they were already testing. In a good number of states, moreover, policymakers did not believe grade-by-grade testing was necessary or desirable. Why would they all move to an annual testing system and how in the world would they pull it off?

What I hadn't considered was the confluence of events that would lead to the reauthorization of the Elementary and Secondary Education Act: a Republican president who believes in testing and accountability from a state that has shown that grade-by-grade testing can help raise achievement; his ability to get key members of his own party in Congress to stop viewing state standards and tests as an intrusion in local control of schools but rather a lever to improve them; and the leadership of key Congressional Democrats, who have come to see the power of standards and tests as a tool for achieving greater equity in American education and improving the life chances of the poorest children.

Now that the legislation has passed and the bill has been signed by the president, the question remains: how are states going to pull it off? The new ESEA amendments require states to begin administering annual tests in grades 3 through 8 in reading and math by the 2005-2006 school year. The previous law required states to test all students in those subjects but only twice within that 6-grade span. Only 16 states currently have grade-by-grade tests in reading and math, and only 9 of those states have tests aligned with their standards (a requirement of the law). The rest will have to fill in the blanks with new tests. Achieve estimates that well over 200 new state-level tests will have to be created over the next several years to meet the new federal requirements.

3 Big Questions

States have made great progress over the last ten years in setting academic standards for students and communicating those expectations to schools and parents. Most states have also tried to align their assessment systems with their standards so that what they are testing becomes more transparent for educators and parents and so that whatever "stakes" are attached to the test results are matched by reasonable opportunities for children actually to learn that which they're being held responsible for knowing. There is still considerable room for improvement, to be sure. But the groundwork is in place in nearly every state. As states move forward to fill in the gaps in their annual testing system, it is

BEST COPY AVAILABLE

critical that the quality of the new tests and their alignment with state standards not get sacrificed.

Are States Ready?
Are states ready to respond to this challenge? It's too soon to be sure. Some states already have tests in all but one or two grade levels, so they only have to create a few new tests. But most states will have to more than double the number of tests they are now giving, and in doing so they will face both educational and political challenges (and incur financial costs as well). The educational challenges have to do with the quality of the tests and their usefulness in improving teaching and learning. This is something that states are already struggling with. The political challenges involve state and local control tensions and sustaining support from educators, parents, and business and community leaders.

Optimally, states will view the federal legislation as an opportunity to take a fresh look at their standards, assessments, and accountability systems and do what it takes to strengthen them. The goal should not simply be to fill in the blank years with tests so that every student is being tested in every grade. Rather, the goal should be to intelligently craft an assessment system that provides teachers, schools, and parents with the data they need to focus attention and resources and achieve better results.

Is the Market Ready?
Directly related to the question of state capacity is the capacity of the testing industry. One of education's dirty little secrets (made less secret last spring by a series of investigative reports by *The New York Times*) is that four major publishing companies have a virtual monopoly on the state testing market. While a few smaller firms have made some inroads over the last several years, the "big four" dominate this $700 million a year industry, creating and administering the tests in most states.

This raises some urgent questions: do these few companies have the capacity to develop over 200 new tests in a very short period of time? The normal cycle for creating a new assessment *in just one state* is 2-3 years. This now needs to happen in two subject areas in multiple grade levels *in at least 34 states*! In order to meet this demand, will the companies be forced to sacrifice their own (variable) standards of quality? Will they end up recycling old test questions and putting together hasty processes for creating new questions, thereby lowering the quality and sophistication of the assessments?

Is the Public Ready?
No matter how states approach the development of their new assessments, their greatest challenge by far will be sustaining the support of educators, parents, and the broader public as the new tests and accountability measures get rolled out. In poll after poll, parents, voters, taxpayers and opinion leaders have said they support testing, even high-stakes testing, because it provides them with some assurance that schools are effectively teaching and students are successfully learning. Educators have been less staunch in their support. They generally agree with raising academic standards, and acknowledge that

BEST COPY AVAILABLE

tests are needed to measure achievement, but their support has begun to waver as real accountability measures have been put in place.

State and local policymakers will need to be mindful of this as they contemplate how to fill in the gaps in their testing programs. Few educators relish the idea of adding more tests on top of those they already have. States will need to be strategic: as new state tests get added, duplicative local tests should be taken away. And educators are sure to pay attention to what the new tests are measuring. The narrower and less sophisticated the questions, the more we will hear complaints from teachers that they are being forced to water down—or narrow—their teaching and focus on a test-prep curriculum.

The Challenge Ahead

At its core, the new law challenges states to measure student achievement more often in order to ensure that students are progressing on a path to proficiency. The idea is not to wait several years before taking the students' academic temperature, but rather to do it in every grade. More frequent testing leads to more frequent feedback to teachers, students and parents. And that feedback should allow schools to focus instruction where it is most needed and address achievement gaps for the benefit of all students. It is also intended to enable policy makers to intervene in situations where the testing reveals inadequate progress being made.

There are, however, a number of challenges to making this work as conceived, and although the law lists some important criteria state assessments will need to meet, Congress has left many of the toughest decisions to the U.S Department of Education and to the states themselves.

As states fill in the gaps in their testing systems, here are some of the things to watch out for: Will the new tests be adequately aligned to state standards? How challenging are those standards--are they worth aligning to? Will the new tests be aligned with existing tests, such that they measure a logical progression of skills from 3rd to 4th grade, from 4th to 5th and so on through 8th grade? Will the tests be sufficiently challenging? Will they measure advanced concepts as well as basic skills? Will the results be comparable across school districts within each state? How rigorous an approach will each state take to defining what it means to be "proficient"? How quickly and effectively will states report scores back to schools and households? Will states be mindful of the testing burden and work with districts to ensure that, as new tests get created, old ones head for retirement?

The governors, business, and education leaders who attended the 2001 National Education Summit last fall anticipated many of these issues and committed themselves to a set of principles that, if followed, will lead to stronger assessment and accountability systems. States that successfully address these challenges will end up taking maximum advantage of the opportunities the new law affords. Those that do not may very well end up taking a step backward in their reforms.

BEST COPY AVAILABLE

Testing Principles adopted at 2001 Summit:

- *Quality* – State tests should be designed to measure student progress against clear and rigorous standards. Reports sent to schools and parents should indicate how students perform against the standards — not just how they compare with other students. Tests developed for other purposes cannot meet this need. The tests should measure the full range of knowledge and skills called for by the standards, from basic to most advanced.

- *Transparency* – In a standards-based system there should be no mystery about what is on the test. Students, parents, and teachers should know what is being tested. They should be confident that if students are taught a curriculum that is aligned with state standards, they will do well on state tests. The best way for states to ensure transparency is to publicly release questions from previous years' tests, along with sample student answers at each performance level.

- *Utility* – Ultimately, it is the clarity of the results and the manner in which they are used that will make a difference in schools. Test results should be returned to schools and parents as quickly as possible without compromising the quality of the test instrument. Score reports should be clear, jargon-free, and designed to guide action.

- *Comparability* – The goal of state assessment programs is to create measurement systems that can accurately track and compare student and school progress from year to year. To accomplish this, the tests from one grade level to another must be aligned with state standards, and the results must be comparable from grade to grade so that student progress can be tracked from year to year.

- *Coherence* – State tests are only one piece of a comprehensive data system. Local and teacher-developed assessments are important too. States must work with districts to ensure that all tests serve a distinct purpose, redundant tests are dropped, and the combined burden of state and local tests remains reasonable.

- *Strategic Use of Data* – Closing the achievement gap can only occur if student achievement data is disaggregated by race and income, and if schools are required to show that all groups of students have made reasonable progress. By regularly reporting how every school is performing against state standards, states can focus attention on the problem, on the progress that some communities and schools are making in response, and on areas where additional work is needed.

BEST COPY AVAILABLE

How Will States Respond? Four Scenarios

While ESEA lays down some clear markers on issues of academic standards, testing, and accountability, states have numerous options in determining how to fulfill the requirements. The Department of Education will either need to get much more concrete about what is expected or the states will end up determining the answers to these questions themselves. It is worth playing out several plausible scenarios to highlight the costs and benefits of the different approaches states might take.

Scenario #1—Cheap and Easy

It is more costly and time consuming to create new tests aligned with state standards than to take existing tests off a publisher's shelf and assert that they are aligned. The fastest, cheapest way for states to fill in the gaps in their testing programs is to purchase ready-made tests such as the Stanford 9, Iowa Test of Basic Skills, and Terra Nova. These are in widespread use in schools today, but they are not designed to measure student attainment of any particular state's standards. Rather, their main purpose is to compare one student's achievement against that of other students in a national sample, in essence comparing that child against an average.

Comparing pupil performance to an average or "norm" is very different than measuring whether or not that child has met a specific set of academic targets. The targets, or standards, provide something for students and teachers to aim for, and those standards do not fluctuate based on how other children are doing.

Although it is not impossible for commercial tests to be well aligned with states' standards, it is highly unlikely. In studies that Achieve has conducted for states, we have found that commercial tests typically touch on some standards but miss the mark on others. The pattern is that commercial tests tend to focus on what is easiest to assess, and it is often the most rigorous knowledge and skills that are not adequately measured. The result is a testing system that is out of sync with what states profess they want students to learn.

If, therefore, states opt to use "off-the-shelf" tests to fill in the grades where they do not currently have tests, they will likely sacrifice the measurement of their standards in those grades. A combination of customized tests in some grades and off-the-shelf tests in others may also end up sending mixed signals to schools and parents about what students are expected to learn. If, for example, a state uses customized tests in 4th and 8th grades and off-the-shelf tests in the other grades, the 4th and 8th grade teachers may end up paying attention to the state standards because that is what is being tested, but the teachers in the other grades may pay less attention to the standards and more attention to what's on the commercial tests. Imagine a school trying to organize its curriculum in such an environment; imagine teachers trying to collaborate across the grades; imagine parents trying to make sense of their children's test scores from grade to grade.

There is a twist on this strategy that a few states have pursued. In order to get a testing system in place quickly, California began in 1998 by adopting a series of off-the-shelf tests for grades 2-11 (the Stanford 9) and then worked with the testing company

BEST COPY AVAILABLE

(Harcourt Educational Measurement) to adapt or "augment" those tests over time to align better with the state's own standards. Starting in 1999, California children began taking the augmented version of the tests, called "STAR" exams (Standardized Testing and Reporting System). These exams consist of a combination of questions from the Stanford 9 and new test questions that were added to reflect the California standards. According to state officials, as many as 75% of the test questions in math had to be created from scratch to align with the standards; a smaller number of new questions were needed in English.

Although education officials in California readily admit that their unorthodox approach caused confusion and even skepticism in schools across the state, they seem optimistic that their transitional strategy will result in tests aligned with their standards. Before other states consider trying this approach, though, it is worth a more careful look: Just how different are the "augmented" tests from the original ones? How well do they in fact align with the state standards (which, by the way, are among the most rigorous in the nation)? If they do, in fact, align well, how much of that has to do with the fact that California's size and market share allowed it to push the testing company harder than a typical state could? Most states find that they have little leverage over these companies, but big states have greater influence due to the size of their student populations and the huge markets that get opened up for textbooks and other products.

The truth is, alignment of tests with standards is difficult to achieve. Even states that have created their own tests from scratch have had a hard time measuring their standards well. But getting it right will be essential if the new assessments that states create are to add value to the existing ones, and become tools that teachers, parents, and policymakers can rely on to raise student achievement. Doing that well is not apt to be cheap.

Scenario #2—Leave it to Districts

As state leaders have pondered how they're going to fill in the grades where they currently do not have tests, some have said that they would rather let districts use their own local tests in the years when the state does not test. This is clearly the most politically convenient solution, as it sidesteps the state/local tensions and allows districts that already test students in grades 3-8 to leave those tests in place. It does, however, raise serious questions about the comparability of data across those districts.

Formal studies by the National Research Council and informal studies by Achieve have concluded that it is nearly impossible to compare results of different tests in any meaningful way. This is because different tests measure different concepts and skills, so proficiency on one test rarely translates to proficiency on another. If states were to pursue this path of least resistance, therefore, they will likely sacrifice the ability to compare achievement results across districts in the grade levels where the state itself does not test. How important is this to states? Will the lack of a common test in each grade skew the accountability system? Which tests will be factored into the adequate yearly progress formula: the state tests, the local tests or both? How can one provide cumulative results for the state as a whole if the tests differ from place to place within it? Wouldn't that lead

BEST COPY AVAILABLE

to data that are very difficult to disaggregate? Will multiple tests send conflicting signals to schools as to where they should focus their curriculum and instruction?

Scenario #3—New Customized Tests

In order to stay true to the principles of alignment, coherence, and comparability, the most desirable strategy for building an annual testing system is for states to develop new tests for the grades where they don't have them. Those tests would be both aligned to the their academic standards and aligned with the tests that they already have.

There are several different ways states might approach this. Some may choose to match the length and sophistication of their existing tests. Other states may decide to alter the format and length of their new tests. They may do this to reduce costs, to reduce the amount of time needed for students to take the tests, or to make the tests more diagnostic and useful to local educators. This is where a creative approach to the task could have the greatest educational payoff.

Imagine a state that currently has reading and math tests in 3rd, 5th, and 8th grades, and each of those tests is 90 minutes long and consists of a combination of multiple-choice and extended response questions (i.e., questions requiring written answers, such as essays). Confident in the data those existing tests provide and wary of the costs of producing identical tests in new grades, state officials might decide to create a shorter version for grades four and seven designed to provide a brief snapshot in between the other tests. The new tests might have fewer questions or rely more heavily on multiple-choice questions, and might only require 45 minutes of test-taking time. This approach would allow states with sophisticated assessments to maintain them at some grades while using more economical versions at other grades.

Another approach might be to make the new tests as sophisticated as the existing tests, but to get creative in how they are scored. Indiana is one state considering this. The idea officials are exploring is to have classroom teachers scoring certain portions of their students' tests and to make the results immediately accessible to schools and parents. There would clearly be quality control and consistency issues that the state would need to work out, but in addition to saving money on centralized scoring, one of the benefits of this approach is that teachers would be much more invested in the assessment process and, therefore, may end up using the results in their classrooms. In fact, done right, grading state assessments could be a very effective form of professional development. Indiana is also exploring the development of formative assessments that teachers can voluntarily use at any point during the school year to determine how their students are advancing toward the state standards.

However states approach the task of creating new tests, it is critical that they remain vigilant about test quality. Achieve's work has revealed that even states that have created their own assessments for the purpose of measuring their own standards have had a difficult time getting it right.

BEST COPY AVAILABLE

Scenario #4—State Collaboration

When it comes to creating high quality tests worth teaching to and basing serious accountability systems on, the deck is clearly stacked against most states. High quality tests cost more to create and there is a limited pool of talent available to help them accomplish this. Given these tensions and the real pressure that states are under to get so many new tests in place relatively quickly, it is legitimate to ask why states need to go it alone.

The most logical strategy for responding to the ESEA testing requirements is for states to pool resources and develop common assessments that they can share. This would allow states that do not have the market power of California, New York, and Texas to work together to leverage better quality tests. They are all relying on the same few companies to create these tests. Why not step back, form strategic partnerships, and leverage the situation?

There are three reasons that states should consider doing this. The upsides are better quality tests, lower costs, and more comparable data across states since they will be using the same tests. The cost savings could be significant at a time when state budgets are tight and it's not clear whether Washington is earmarking enough money to offset state testing costs. The comparability advantage also deserves more attention than it typically gets: one reason the legislation requires all states to give NAEP reading and math assessments every two years is that policymakers want better ways to compare results across states against a common standard. Why not build that comparability into states' own assessment systems while they have the chance? This happens to be the reason some state policymakers and parents like the idea of using norm-referenced tests—it gives them some ability to compare results beyond their state.

The new law specifically allows states to form consortia and pool resources to create and use common tests. The main thing standing in the way at this point seems to be habit. States are used to working individually with test publishers to create their own tests. They are not used to a collaborative approach. This may change as states look ahead at the need to build over 200 new tests.

There is at least one consortium already in place that could be very helpful to states as they develop their ESEA strategies. At the request of governors and education commissioners in a number of states, Achieve launched an initiative in 1999 known as the Mathematics Achievement Partnership to help states work together to raise mathematics standards and achievement. Fourteen states are currently involved in the partnership, which will provide them with an internationally benchmarked 8th grade math assessment, tools for improving the middle school math curriculum, and strategies for improving the professional development of middle school math teachers. We are exploring how states can tap into the consortium to develop tests in the grades where they currently do not have them.

Getting It Right

The task ahead for states in building an annual testing system reminds me of what must be a fairly typical challenge facing city planners when they address changes in traffic patterns. Oftentimes, heavier usage on some roads necessitates adding stop lights at more intersections to control traffic and ensure safety. When confronted with the challenge of adding traffic lights at more intersections along a busy street, what would a thoughtful city planner do? Would he purchase the least expensive product even if the signals it sent were different than those of the existing traffic lights? Would he ask the residents on each block to build or buy their own traffic light? How would traffic be affected if the new signals were not timed with the existing ones? Would it help control the flow of vehicles or simply confuse and frustrate drivers and pedestrians?

The thoughtful city planner keeps the endgame in mind as he devises his plan. The goals are safety and the smooth flow of traffic, not placing a traffic light at each intersection. That's simply a means to the end. If poor decisions are made, it is quite possible that the addition of lights at each corner could make the streets more congested and less safe.

It is the same with building an annual testing system. Approached intelligently, grade-by-grade testing can be a real improvement over what many states currently have in place. But not all strategies for creating annual tests will result in a coherent assessment system. States must take care to get it right.

The President and Congress did make an effort to address some of the issues discussed in this paper. There are a series of criteria laid out in the law that state assessment systems will need to meet. These include: alignment with state standards; reporting scores for each individual student; disaggregating the data by race, ethnicity, and socio-economic status; providing itemized analyses pointing to students' strengths and weaknesses in each particular skill area; returning the results before the beginning for the next school year; and assessing "higher order thinking skills and understanding."

At this stage, the question on most people's minds is how rigorous federal officials will be in their interpretation of these criteria and, more importantly, how serious they will be about enforcing them. Federal officials can and should play an important role in clarifying criteria and reviewing state plans, and if they take a hard line on some of these important issues, states could be left with a smaller but smarter set of options.

If past experience is our guide, however, we should not expect the federal government to fully solve complex issues such as the quality, alignment, comparability, coherence and utility of state standards and assessment systems. The federal government can lay down clearer markers and use the bully pulpit, but in the end, these are issues that state leaders must address for themselves.

Using NAEP to Confirm State Test Results: An Analysis of Issues[1]

Mark D. Reckase
Michigan State University
February 18, 2002

The new Elementary and Secondary Education Act (ESEA) amendments, "No Child Left Behind," require that the National Assessment of Educational Progress (NAEP) reading and mathematics tests be administered every other year in grades 4 and 8. Further, states must participate in the component of NAEP that is used to obtain estimates of students' academic performance at the state level. This part of the NAEP program is called State-NAEP. Participation in State-NAEP has been voluntary in the past, but the ESEA amendments make participation a condition of accepting Federal funds related to the legislation. While the legislation does not indicate what is to be done with the results of NAEP testing, it does imply that NAEP will be used as a check on the reading and mathematics assessment results reported by each state. Further, states will be required to administer their own reading and mathematics assessments to their students every year in grades 3 through 8. The purpose of this policy memo is to summarize the issues related to the use of NAEP to confirm the assessment results reported by states.

Testing Programs in the ESEA Legislation

A Brief Description of NAEP

NAEP is an extensive program of data collection that includes achievement tests in a number of subjects, including, but not limited to mathematics and reading. NAEP also collects information about characteristics of the student population and features of the educational system. NAEP results, and the many interpretive reports produced from those results, provide an ongoing description of the functioning of the educational systems in the United States.[1]

NAEP tests are uniquely different from state assessments in a number of ways. First, the tests attempt to measure student capabilities (what students know and can do) on a domain of process and content knowledge that is common to the state educational systems across the United States. The creators of the document describing what is included in that domain also attempt to include content and processes recommended in future-oriented standards documents (e.g., those promulgated by the National Council of Teachers of Mathematics) so that the domain definition will be applicable for a number of years into the future. Allowing the national standards documents to influence the domain definitions implies that states' are expected to move their curriculum in the direction of those standards.

---

[1] Paper prepared for "Will No Child Truly be Left Behind? The Challenge of Making this New Law Work – a conference sponsored by the Thomas B. Fordham Foundation, Washington DC, February 13, 2002.

The domain of coverage for a NAEP subject matter area is described in a document called a "framework" (e.g., *Reading Framework for the 1992 National Assessment of Educational Progress* (NAGB (1992)). A consequence of the need for NAEP to be appropriate for assessing student performance in all states is that it can not focus too closely on the educational goals from any one state. NAEP assesses the common core of all state programs, but it does not assess the instructional goals that are unique to individual states.

A second way that NAEP is unique is that no student takes the entire test. Because NAEP endeavors to assess what students know and can do in a very broad domain, the full NAEP tests contains a large number of questions --145 to 160 questions for NAEP Mathematics, for instance. This number of questions is too large for any student to attempt in a reasonable period of time. To keep thorough domain coverage, but also keep the testing time to a reasonable amount, each student takes only 36 to 45 mathematics questions. Test booklets contain overlapping sets of questions so that the results from all of the examinees can be combined to determine the expected distribution of performance on the full set of questions for the full sample of students. However, it is not possible to obtain a good estimate of performance on the full domain of knowledge and skills for any individual student because the student has responded to only a small part of the entire test.

A third unique feature of NAEP is a direct result of the item and student sampling approach that it uses to keep testing demands within reasonable bounds. Because students take only part of the test, no student scores are reported. Also, tests are only administered to a random sample of students from the nation and from within participating states. A consequence of the sampling approach is that only estimated score distributions for state and national groups can be reported. NAEP summarizes the information from these distributions using percentages above achievement levels set by the National Assessment Governing Board (NAGB) and descriptive statistics (means and standard deviations). It is not possible to track individual student's performance on NAEP over years or directly compare student performance on NAEP with that student's performance on a state test. Nor is it possible to report NAEP results at the school building level because only a small number of students from any school take the test, and those students take only part of the full set of test questions.

The unique features of NAEP have not interfered with its use as a general indicator of the quality of education in the United States. However, they will need to be taken into account when NAEP results are compared to state results.

State Assessments

State assessment procedures are notable for the diversity of approaches that they take. Some states purchase existing tests from commercial test publishers as all or part of the state assessment program. This approach would seem to indicate that these state education officials believe that the commercial tests are sufficiently aligned with the curriculum and instruction goals for the state. Other states hire test development

contractors to custom develop elaborate assessment programs according to state developed test specifications. The test specifications for these programs vary greatly. Some include performance assessment tasks that are scored by commercial companies, others are multiple-choice only, and some use computerized testing procedures as part of the assessment program. One state (Iowa) does not have a state assessment program, though most students in the state take the Iowa Tests of Basic Skills and Iowa Tests of Educational Development at some point in their schooling.[2]

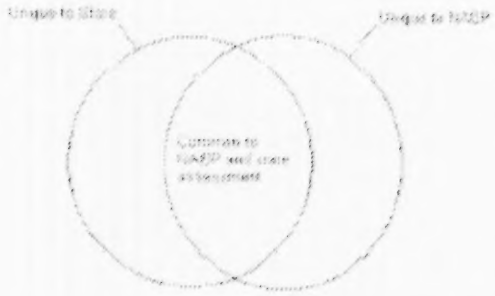
The diversity of state assessment programs provides a challenge for the use of NAEP to confirm the results of those assessments. The state assessment programs have different content, schedules for administration, purposes, stakes, and technical characteristics. Further, many of these features will likely change in response to the ESEA legislation. At the very least, many states will have to increase the frequency of testing in grades three through eight in reading and mathematics. The next section of this memo highlights a number of the more important issues related to the use of NAEP for confirmation purposes. The following sections discuss the effects of differences in state testing programs on the interpretation of NAEP/state assessment comparisons.

The Relationship between NAEP and a State Assessment

Domain Overlap

The starting point in the design of an achievement test is the specification of the domain of content and skills to be covered by the test. In theory, there should be a description of the domain at a level of detail that will allow an interested party to determine whether a specific test task should be on the test because it measures part of the domain, or whether it should be excluded because it does not. The NAEP framework documents are good examples of domain specifications. Unfortunately, like everything else with state assessment programs, descriptions of domains vary substantially across states. Some give very general statements of academic goals; others provide detailed descriptions of desired academic content and skills.

A key to determining the comparability of NAEP and state assessment results is an evaluation of the commonality of the target domains. The following diagram gives a simplified representation of the overlap in those domains. The content domain for a state assessment program is represented by one circle and the domain for NAEP is represented by another circle. Within a circle is the content and skills to be measured by an assessment program. Outside the circles are the content and skills that are not included in the domains for either of the two assessment programs. For each assessment, there is part of the domain that is in common with the other assessment and part that is not.



States vary in the amount that their assessment domains overlap with NAEP. For some, there is almost complete overlap. For others, the overlap is modest. Unfortunately, there do not seem to be any formal studies of the amount of overlap between domains for NAEP and state assessments. Such studies would be major undertakings that would require in-depth analysis of every state testing program. There would be a further complication that state assessment programs are not static – they change frequently, sometimes because of changes to the curriculum, but also because of other factors such as the need to reduce costs, or because of changes in educational policy within the state.

Assuming that the amount and composition of domain overlap can be determined, a critical issue when comparing NAEP and state assessment results is the part of the domain that is emphasized by instruction within the state. Although a state may define a large content and skill domain as the focus of instruction, not all parts of that domain will be treated with the same emphasis in every classroom. If the focus of classroom instruction is on parts of the state's domain that do not overlap with the NAEP domain, then student performance may improve and be documented on the state assessment while that improvement is not shown on NAEP. NAEP might even show a decline if the part of the domain that is common to the assessment programs and the part that is unique to NAEP are given little instructional emphasis.

To the extent that NAEP has captured the important outcomes of the nation's educational systems, the cases of low domain overlap and of instructional focus on things not covered by NAEP should be rare. But it is possible that a state could show improvement, NAEP could show decline, and they could both be correct because instruction is focusing on different parts of the combined domain for the two tests.

Performance Standards

BEST COPY AVAILABLE

NAEP reports results in two ways. The first is estimated test score distributions on the NAEP standard score scale. This type of reporting includes mean scores for demographic groups and state samples. The second way that NAEP results are reported is percentages above achievement levels set by NAEP's governing body, the National Assessment Governing Board (NAGB). NAGB has set three such levels labeled basic, proficient, and advanced. The achievement levels are ranges between cut scores on the NAEP score scale. NAGB considers these cut scores as definitions of performance goals for what students should know and be able to do at grades 4, 8, and 12.[3] The NAGB achievement levels take on special meaning in the ESEA legislation because the legislation specifies that states must define their own "proficient" and "advanced" levels, as well as a "basic" level. The language of the legislation uses the same labels already used for the NAGB-developed achievement levels on NAEP.

States also set cut scores on their assessments, but even when they use the same labels as the NAGB achievement levels the meanings of the state standards might be quite different. For example, a state may use the term "proficient," but in terms of the number of students who attain that level or higher, the state's proficient level may be similar to the NAGB "basic" level. Such differences in meaning of state and NAGB standards are not likely a sign of duplicity. The research on standard setting shows that different standard setting methods, different statements of policy, and standard setting panels with different characteristics are likely to produce different standards.[4]

The location of cut scores on a score scale is important because the location indicates where the reporting system will be sensitive to changes in student performance. Consider the following thought experiment. Suppose that a standard is set on a mathematics test by placing a cut score for reporting at roughly the level of difficulty of simple addition problems. Also suppose that at grade 4 in one school, the students are not yet doing well on addition, while at another school most of the students have mastered addition. In the first school, if instruction focuses on simple addition, many students will move from below the standard to above the standard. It is likely that the percent above the standard will improve quite dramatically. In the second school, however, because the students already know the material and because instruction is focused on other, probably higher level skills and knowledge (e.g., fractions), the increase in percent of students attaining that state's standard in that school will be small. The opposite effect can occur if the cut score is set at a level that is consistent with the difficulty of the fraction problems. In that case, the second school would show a lot of improvement and the first school would show very little.

The NAGB "proficient" level is a fairly high standard. Changes in the percent above that standard will likely reflect achievement gains for students whose instruction focuses on the more difficult NAEP content. Changes in the proportion above "basic" will likely show improvements for students whose instruction focuses on relatively easy NAEP content.

Context of the Assessment

Not only do NAEP and state assessments differ on domain coverage and the placement of performance standards, they also differ in the context for the assessment; that is, the way that the assessment is perceived by the students and the local school district staff. For example, some states use their assessments to determine whether students will be promoted to the next grade or whether school staff will receive monetary awards for helping students reach instructional goals. These assessment programs are called "high stakes" because there is a direct and important consequence to the students and school staff. In such cases, it is likely that students will be motivated to do well and the school staff will do what they can to help the students perform at their best.

The amount of "stakes" for state assessments varies quite dramatically. Some states use the assessment results only for general school accountability purposes with no direct consequences for students. Some states test a sampling of students rather than every student. Other states make the assessments a very important part of the state instructional system. Teacher salaries may depend on the assessment results and students may receive direct rewards or punishments. The high level of variability across states with regard to "stakes" adds to the complexity of comparing state results with NAEP results.

NAEP has no direct consequences for students or school staff because NAEP results are not reported at the school or student level. Students do not receive scores and schools do not receive summaries of student performance. These features of NAEP make it a "low stakes" assessment at the school and student level. The differences between contexts for state assessments and NAEP need to be taken into account when interpreting comparative results.

Analysis

When comparing state assessment results with NAEP results for a single curriculum area, there are nine possible results as depicted in the cells in the following table. NAEP confirming state results would seem to require that both testing programs have results in the cells with the Xs. The question of concern here is "How likely is it that NAEP and state assessments will give results in these cells?" To answer this question, all of the issues that have been summarized need to be considered.

		State Assessment		
		Decline	Stay Level	Increase
NAEP	Decline	X		
	Stay Level		X	
	Increase			X

First, the issue of domain overlap needs to be considered. For most states, the domain overlap between NAEP and the state assessment will be at least moderate. NAEP was designed to measure the common content of the instructional systems of all of

the states. Unless a state has instructional goals that are notably different than those of other states, there should be some commonality between domains of coverage for NAEP and a state assessment. However, it is not likely that the overlap will be total for any state. It is possible that there may be important parts of a state domain that are unique to the state and not included in the content of NAEP. If the state focuses instruction and assessment on the unique features to the exclusion of the common components, it is possible for the state assessment to show gains when NAEP does not. It is also possible for NAEP to show gains when a state assessment does not if instruction focuses on the unique features of NAEP (e.g., instruction may be focused on national curriculum standards) rather than the unique features of the state assessment. This seems less likely, but possible. The existence of these possibilities suggests that part of the interpretation of NAEP results for confirming state results will need to be a judgment of the overlap between the assessment domains. Substantial overlap makes NAEP a stronger tool for confirmation. Low overlap indicates that NAEP can not provide solid evidence for confirmation or disconfirmation.

Second, the context of the state assessment will also likely affect the usefulness of NAEP as a source of evidence for confirmation. If the state assessment is high stakes and NAEP is low stakes, students may try very hard on the state assessment and not very hard on the NAEP. Real situations may be more complicated. There are more possibilities than motivated and not motivated. Students vary in level of motivation and the level of student motivation may interact with the level of difficulty of items. Students may give a reasonable level of effort to easy items even when the test does not count for them, but they may give up on hard items when the test does not have direct consequences. The result of differences in stakes may be that students show improvement on the state assessment if it is high stakes and no improvement or a decline on NAEP.

The context of state assessments and NAEP may differ in other ways that may affect the comparison of results. The assessment programs may be administered at different times of the year. If the state assessment is administered in the fall, and NAEP is administered in the spring, the amount of exposure to the curriculum will differ. The differences in instructional time will influence the amount that students have learned by the time the test is administered and the amount of gain that can be detected. The quality of the assessments may also differ, affecting the confidence that can be placed in the reported results.

The location of standards on the assessment can result in similar differences in results. Students at all points in a distribution of performance will not likely improve by equal amounts. If a school focuses on the improvement of basic skills, performance standards set at a relatively low level will show the greatest change in the percent attaining those standards. The NAGB "proficient" level is a high standard so it may not be sensitive to changes in basic skills. A basic skills oriented state standard might show improvement while the percent above NAGB "proficient" does not. The opposite may occur for schools focusing instruction at a higher level – NAEP may show changes when the state assessment does not.

A solution to this problem is to look at changes at all levels of student achievement rather than at single cut scores. NAGB is currently investigating reporting procedures for NAEP that can show changes along the entire NAEP score scale. These same procedures could be used by states as well.

The description of state and NAEP assessment programs given here is based on the current characteristics of those programs. However, the legislation will likely result in significant changes to both NAEP and state assessments. A recent review of state testing programs in *Education Week* indicates that only eight states currently meet the requirements set out in the legislation. Many states will have to expand their reading and mathematics assessments to meet the requirement of testing every year from grade 3 to grade 8. NAEP will also have to change its testing schedule to provide results every other year in mathematics and reading. While it is likely that significant changes in these assessment programs will occur, the full impact of the changes will not likely be understood for several years.

Conclusions

Jointly interpreting state assessment and NAEP results in a coherent way will not be a simple task. Many factors need to be taken into account when making such interpretations including the amount of content overlap, the location of cut scores on the score scales, and the context for the assessments. This is not to suggest that the joint interpretation of the test data is impossible or unwise. Experience from analysis of ACT and SAT college admissions tests and other testing programs indicates that tests constructed from different test specifications can yield highly correlated results. It is likely that NAEP results and state assessment results will be related as well. With careful consideration of threats to accurate interpretations and realistic judgments about the amount of effort that will be required to make accurate interpretations, joint use of NAEP and state assessment results should lead to better understandings of the functioning of the educational systems in the United States.

References

Bourque, M. L. & Byrd, S. (Eds.) (2000). *Student performance standards on the National assessment of educational progress: affirmation and improvements.* Washington, DC: National Assessment Governing Board.

Braswell, J. S., Lutkus, A. D., Grigg, W. S., Santapau, S. L., Tay-Lim, B. & Johnson, M. (2001). *The Nation's Report Card: Mathematics 2000 (NCES 2001-517).* Washington, DC: National Center for Educational Statistics.

Cizek, G. J. (Ed.) (2001). *Setting performance standards: concepts, methods, and perspectives.* Mahwah, NJ: Lawrence Erlbaum Associates.

Meyer, L., Orlofsky, G. F., Skinner, R. A. & Spicer, S. (2002). The state of the states. *Education Week, 21*(17), 68-169.

National Assessment Governing Board (1992). *Reading framework for the National Assessment of Educational Progress.* Washington, DC: Author.

---

[1] Details of features of NAEP are presented in a number of documents including Braswell, Lutkus, Grigg, Santapau, Tay-Lim and Johnson (2001).
[2] A brief summary of state assessment programs is given in Meyer, Orlofsky, Skinner and Spicer (2002).
[3] For a discussion of the issues related to the standards set by NAGB, see Bourque and Byrd (2000).
[4] See Cizek (2001) for recent information on standard setting.

U.S. Department of Education
Office of Educational Research and Improvement
(OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

Reproduction Release

(Specific Document)

TM035280

I. DOCUMENT IDENTIFICATION:

Title: Multiple Choices	
Author(s): Matthew Gardal Achieve, Inc.	
Corporate Source:	Publication Date: February 2003

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, Resources in Education (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign in the indicated space following.

The sample sticker shown below will be affixed to all Level 1 documents	The sample sticker shown below will be affixed to all Level 2A documents	The sample sticker shown below will be affixed to all Level 2B documents
PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY SAMPLE TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY SAMPLE TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY SAMPLE TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
Level 1	**Level 2A**	**Level 2B**
☑	☐	☐
Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g. electronic) *and* paper copy.	Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only	Check here for Level 2B release, permitting reproduction and dissemination in microfiche only
Documents will be processed as indicated provided reproduction quality permits. If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.		

ACHIVE, INC. Fax:2026243671 Oct 21 2003 14:07 P.01

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche, or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Signature:	Printed Name/Position/Title: TIFFANY LYN PACHE, Research Analyst	
Organization/Address: 400 N. Capitol Street, NW Suite 351 Washington, DC 20001	Telephone: 202-624-1400	Fax: 202-624-x1408
	E-mail Address: tpache@acheive.org	Date: Oct. 21, 2003

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, or, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:
Address:
Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:
Address:

V. WHERE TO SEND THIS FORM:

ARCHIVE, INC. Fax:2026243671 Oct 21 2003 14:08 P.02